DOCUMENT RESUME

ED 385 561 TM 023 983

AUTHOR	Freedle, Roy; Kostin, Irene
TITLE	The Prediction of GRE Reading Comprehension Item Difficulty for Expository Prose Passages for Each of Three Item Types: Main Ideas, Inferences and Explicit Statements. GRE Board Professional Report No. 87-10P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-91-59
PUB DATE	Apr 92
NOTE	57p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Correlation; *Difficulty Level; Higher Education; Inferences; Multiple Choice Tests; *Prediction; *Reading Comprehension; Regression (Statistics); *Test Items
IDENTIFIERS	Explicitness; *Graduate Record Examinations; Main Idea; Scholastic Aptitude Test

ABSTRACT

This study examines the predictability of Graduate Record Examinations (GRE) reading item difficulty (equated delta) for the three major reading item types: main idea, inference, and explicit statement items. Each item type is analyzed separately, using 110 GRE reading passages and their associated 244 reading items; selective analyses of 285 Scholastic Aptitude Test (SAT) reading items are also presented. Stepwise regression analyses indicated that the percentage of GRE delta variance accounted for varied from 20% to 52% depending on the item type. Details of item predictability were explored by evaluating several hypotheses. Results indicated that: (1) multiple-choice reading items are sensitive to variables similar to those reported in the experimental literature on comprehension; (2) many of these variables provide independent predictive information in regression analyses; and (3) substantial agreement between GRE and SAT reading predictability was found. Five tables in the text and an additional table in the appendix (correlation between item difficulty and predictor variables) present study findings. (Contains 37 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

GRE®
RESEARCH

ED 385 561

The Prediction of GRE Reading Comprehension Item Difficulty for Expository Prose Passages for each of Three Item Types: Main Ideas, Inferences and Explicit Statements

Roy Freedle
and
Irene Kostin

April 1992

GRE Board Professional Report No. 87-10P
ETS Research Report 91-59

Educational Testing Service, Princeton, New Jersey

BEST COPY AVAILABLE

The Prediction of GRE Reading Comprehension Item Difficulty for Expository Prose Passages for each of Three Item Types: Main Ideas, Inferences and Explicit Statements

Roy Freedle
and
Irene Kostin

GRE Board Report No. 87-10P

April 1992

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Graduate Record Examinations and Educational Testing Service are U.S. registered trademarks of Educational Testing Service; GRE, ETS, and the ETS logo design are registered in the U.S.A. and in many other countries.

Copyright © 1992 by Educational Testing Service. All rights reserved.

Acknowledgments

We want to thank Larry Frase, Grant Henning, and Gordon Hale for helpful comments on earlier drafts of this paper. This work was funded by the Research Committee of the Graduate Record Examinations Program.

Abstract

This study examines the predictability of GRE reading item difficulty (equated delta) for three major reading item types: main idea, inference and explicit statement items. Each item type is analyzed separately, using 110 GRE reading passages and their associated 244 reading items; selective analyses of 285 SAT reading items are also presented. Stepwise regression analyses indicates that the percentage of GRE delta variance accounted for varied from 20% to 52% depending upon the item type.

Details of item predictability were explored by evaluating several hypotheses. Results indicated that (1) multiple-choice reading items are sensitive to variables similar to those reported in the experimental literature on comprehension, (2) many of these variables provide independent predictive information in regression analyses, and (3) substantial agreement between GRE and SAT reading predictability was found.

Introduction

Purpose of Current Study

The primary purpose of the current study is to predict reading item difficulty for each of three GRE reading item types--main ideas, inferences and explicit statement items--which together constitute about 75% of the reading items. To achieve this goal we need to identify a set of variables that earlier studies suggest should be predictive of comprehension difficulty. By obtaining confirming evidence that these earlier identified variables are in fact predictive of GRE reading comprehension item difficulty, this can be taken as evidence favoring the claim that the GRE reading section is in fact a measure of passage comprehension. Such an outcome might lead to modifications in statements made recently by Royer (1990) as well as by Katz, Lautenschlager, Blackburn, and Harris (1990), who have argued that multiple-choice reading tests are primarily tests of reasoning rather than passage comprehension per se--these arguments are presented in greater detail below.

Background Studies

Only a few studies appear to have focused on predicting item difficulty using items from standardized ability tests (Drum, Calfee, & Cook, 1981; Embretson & Wetzel, 1987). While not specifically focused on predicting reading item difficulty, many other studies of language processing have isolated a wide variety of variables that influence comprehension difficulty with respect to decision time and recall measures. A few such studies of particular interest here are the study of negations by Carpenter and Just (1975), the study of rhetorical structure (Grimes, 1975) and its effect on accuracy of prose recall (Meyer, 1975; Meyer & Freedle, 1984) and prose comprehension (Hare, Rabinowitz, & Schieble, 1989); the use of referential expressions in constructing meaning (Clark & Haviland, 1977), and the use of syntactic "frontings" (see details below) that appear to guide the interpretations of semantic relationships within and across paragraphs (see Freedle, Fine, & Fellbaum, 1981). The particular manner in which these selected variables will be employed will become evident later in this report. Using this set of hypothetically relevant variables, the primary strategy employed in this work has been to try to capture the large- and small-scale structures of the reading passages, and their associated items, in order to best account for observed reading item difficulty in a multiple-choice testing context.

First we review those studies that predict reading item difficulty for multiple-choice tests.

Drum, Calfee, and Cook (1981) predicted item difficulty using various surface structure variables and word frequency measures for the text, and several item variables that also depended on surface structure characteristics (e.g., number of words in the stem and options, number of words with more than one syllable). They reported good predictability using these simple surface variables; on average, they indicated that about 70% of the variance of multiple-choice reading item difficulty was explained.[1]

Embretson and Wetzel (1987) also studied the predictability of 75 reading item difficulties using a few of the surface variables studied by Drum et al. (1981). But in addition, because of the brevity of their passages, Embretson and Wetzel were able to do a propositional analysis (see Kintsch & van Dijk, 1978) and add variables from this analysis, along with several other measures, as predictor variables. In particular they found that connective propositions were significant predictors. We believe that Meyer's (1975) top-level rhetorical structures, which we include in the present study, indirectly assess the presence of connectives (such as and, but, however, since, because, etc.) since each of the rhetorical devices differently emphasizes these connectives. For example, a top-level causal structure tends to use connectives such as since and because. A list structure tends to use connectives such as and and then, while a comparative structure will often employ connectives such as however, or yet, etc.

Now we review those additional studies that deal with variables that have been found to influence reading comprehension difficulty. Most of these additional variables were investigated in empirical studies that did not use multiple-choice methods to yield an index of comprehension difficulty. Instead many used dependent measures such as recall of passages or decision time to infer the influence that certain variables have on comprehension difficulty. This review along with our earlier review of the Drum et al. (1981) and Embretson & Wetzel (1987) studies will help us to select a final set of variables that we postulate may also index comprehension difficulty within a multiple-choice testing format.

Carpenter and Just (1975) found that the occurrence of sentence negation increases comprehension decision time. This suggests that the number of negations contained in GRE reading passages may also influence multiple-choice item difficulty. Furthermore, one can inquire whether additional negations that are used in the item structure itself (either in the item stem or among the response options) may also separately contribute to comprehension difficulty over and above the contribution of text negations.

Abrahamsen and Shelton (1989) demonstrated improved comprehension of texts that were modified, in part, so that full noun phrases were substituted in place of referential expressions. This suggests that texts with many referential expressions may be more difficult than ones with few referential expressions. Again, for purposes of studying more broadly the effect of number of referential expressions on comprehension difficulty of multiple-choice tests, a separate count is also made of referential expressions that occur in the item proper.

Hare et al. (1989) studied, in part, the effect of four of Grimes' (1975) rhetorical organizers on difficulty of identifying the main idea of passages--students either wrote out the main idea if it was not explicitly stated or underlined it if it was explicitly stated. They found a significant effect of rhetorical organization such that list type structures (see definitions and examples below) facilitated main idea identification whereas some nonlist organizers made main idea information more difficult to locate. Meyer and Freedle (1984) examined the effect of the Grimes organizers on the ability of students to recall passages that contained the same semantic information except for their top level rhetorical organization. They found, like Hare et al., that list structures facilitated recall (for older subjects). However, they also reported that university students were best helped by comparative type organizations; this latter finding was not replicated by Hare et al.

It seems likely that rhetorical organization will contribute to comprehension difficulty within a multiple-choice testing format; however, it is not clear, given the differences between the Meyer and Freedle (1984) and Hare et al. (1989) studies, whether we can say in advance which type structure will be found to facilitate performance. Top-level rhetorical structure meaningfully applies only to the text structure; a comparable entry for items is not feasible.

Freedle, Fine, and Fellbaum (1981) report differences in the use of "fronted" structures at sentence beginnings (and paragraph beginnings) as a function of the judged quality of student essays. Fronted structures included the following: (1) cleft structures ("It is true that she found the dog," where the initial "it" is a dummy variable having no referent); (2) marked topics consisting of several subtypes (a) opening prepositional phrases or adverbials ("In the dark, all is uncertain"; "Quickly, near the lodge, the boat overturned") or (b) initial subordinate clauses ("Whenever the car stalled, John would sweat"); and (3) combinations of coordinators and marked topics or cleft structures that begin independent clauses ("But, briefly, this didn't stop him"; "And, furthermore, it seems that is all one should say").

Freedle et al. (1981) showed that these different fronting structures significantly discriminate among essay quality such that the better essays contained a higher mean frequency of each of these fronted structures even after partialling out the effect of different lengths of essay as a function of ability level. They interpreted these fronted structures as authors' explicit markers for guiding readers to uncover the relationships that exist among independent clauses. It is not immediately clear whether differential use of all such structures would itself facilitate or inhibit comprehension of GRE passages. If we assume that the structures produced by the more able writers are structures that are more difficult to learn, then one can predict that the more frequently these fronted structures occur, the more difficult the text should be to understand. In support of this, Clark and Haviland (1977) suggest that at least cleft structures may be harder to understand than simple declarative sentences. Also, Bever and Townsend (1979) found that when main clauses follow a subordinate clause, such sentences are more difficult to process than when main clauses occur in initial sentence positions (this overlaps somewhat with frontings, since initial subordinate clauses would count as one type of fronting). by including a count of all such variables, we can explicitly test the relevance of clefts and other fronted structures for how they might affect comprehension difficulty in a multiple-choice testing context. This is done separately for text as well as item content.

While Kieras (1985) specifically focused on the perception of main idea information in reading, his study will be seen as potentially relevant for all three item types treated in this the current study. First we summarize Kieras' earlier work and then generalize it to include inference and explicit statement items.

Kieras (1985) examined, in part, how students perceived the relative location of main idea information in short paragraphs. He found, using single paragraph passages extracted from technical manuals, that most students perceived main idea information as located early in the paragraph and a few thought the main idea occurred at or near the end of the paragraph; information in the middle of the paragraph was least often perceived as a statement of the main idea. Kieras did not report the relative frequencies with which the actual main ideas occurred among the passages so it is difficult to determine whether students tended to select the opening sentences of passages as containing the main idea because most of the passages placed the key idea in this place or whether the students were simply reflecting a response bias to choose the opening sentences. Unless the main idea was equally represented by its location across the stimulus passages, the Kieras results are ambiguous.

However, the work of Hare et al. (1989) helps to clarify this issue. In one of their studies they systematically varied the known location of a main idea sentence in three locations: the opening sentence, the medial sentence, or the final sentence of a paragraph. The experimental subjects underlined which sentence they thought was the main idea sentence. Correct identifications were greatest for initial occurrence of main idea sentences. One can infer from the Hare et al. results that two tendencies contribute to main idea correctness: opening sentences that do contain the main idea tend to be selected partly because of a prior bias to select early sentences, but also because students are attempting to understand the information in the text sentences.

One can generalize the Hare et al. (1989) work and the Kieras' (1985) findings to demonstrate the possible relevance of locational effects concerning how students respond to multiple-choice items for multiparagraph passages. If students tend to perceive early text information, especially information in the opening sentences of the first paragraph, as main idea information (see Kintsch's, 1974, earlier work in this regard), then when certain passages actually confirm this search strategy, such items should be easier than those that disconfirm it (where disconfirming main idea information would be information that occurs in the middle of a multi-paragraph text; it is disconfirming only because it fails to conform to the expectation that main idea information "should" be near the beginning of a passage). So, the relative ordering of difficulty should be: opening sentences that fit the main idea information as stated in the correct answer to a main idea item will be easiest (other things being equal), while main idea information that occurs near the middle of a text will be associated with the hardest main idea items. As we have already suggested, such a result might be expected based on the earlier text processing theory of Kintsch.

Since we also intend to study inference as well as explicit statement items, we might inquire whether the Kieras (1985) and Hare et al (1989) type findings about relative location of information in the passage for main ideas will also help account for item difficulty associated with these other two reading item types.

Explicit statement items are of the following type: "According to the passage, x occurs when...." It seems reasonable to expect that if the relevant explicit information occurs early in the passage, the item should tend to be easy. But if the relevant explicit information is located near the middle of the passage, the item should be more difficult. If so, this generalizes our interpretation of Kieras' (1985) results for main ideas to explicit statement information. We hypothesize that the surface location of

relevant information influences the results. While for main ideas one normally expects early text information to contain the relevant main idea, there is no corresponding expectation for explicit statement information. Nevertheless, the beginning of a passage may be especially salient even for explicit statement items, not because a prior expectation is confirmed or disconfirmed but, more simply, because examinees may start their search for explicit information at the beginning of the passage.

A similar argument can be made for inference type items. Inference items usually have the following format: "One can infer from the passage that x means...." If the relevant text information needed to carry out the inference is located near the beginning of the passage, this might facilitate getting the item correct. But if the relevant text information is in the middle, the item will probably be more difficult.

Other variables that one can hypothesize will be of importance in affecting comprehension difficulty for multiple-choice tests are vocabulary level (Graves, 1986), various measures of sentence complexity, such as sentence length (Klare, 1974-1975), paragraph length (Hites, 1950), number of paragraphs (Freedle, Fine, & Fellbaum, 1981) and abstractness of text (Paivio, 1986). In particular, less frequently occurring words and longer sentence structures tend to make texts more difficult to understand, as can be inferred from their use in traditional readability formulas (Graves, 1986); in addition, longer paragraphs, and abstractness of texts also make passages more difficult to comprehend (see Hites, 1950, and Paivio, 1986, respectively). Use of more paragraphs was found to be positively correlated with the quality of written essays (Freedle, Fine, & Fellbaum, 1981); it remains to be seen whether number of paragraphs itself contributes to reading comprehension difficulty in a multiple-choice testing context.

It will be useful to collect this review of variables expected to influence reading comprehension item difficulty into a single set. Hence one can hypothesize that many of the variables listed that are known to contribute to comprehension difficulty in non-multiple-choice testing formats (or to quality judgments of written essays) will be found to significantly affect comprehension measures as determined within a multiple-choice testing format. Stating this more succinctly we have:

Hypothesis 1. The following variables are expected to significantly influence reading item difficulty as determined within a multiple-choice testing format:

a. negations
b. referentials
c. rhetorical organizers
d. fronted structures
   1. cleft structures
   2. marked topics
   3. combinations (of coordinators and marked topics or coordinators with cleft structures)
e. vocabulary
f. sentence length
g. paragraph length
h. number of paragraphs
i. abstractness of text
j. location of relevant text information

Relevance of hypothesis 1 to criticisms of multiple-choice reading tests as tests of passage comprehension. Hypothesis 1 can be viewed as an important hypothesis, particularly as it applies to the coding of the passage, as we shall endeavor to explain below. Royer (1990) indicates that "There is evidence that standardized reading comprehension tests that utilize multiple-choice questions do not measure the comprehension of a given passage. Instead they seem to measure a reader's world knowledge and his or her ability to reason and think about the contents of a passage" (p. 162). Royer then cites work by Tuinman (1973-1974), Drum et al. (1981) and Johnston (1984) to bolster this claim. Tuinman's work is similar to the findings of Katz et al. (1990) wherein multiple-choice reading items are correctly responded to above chance levels in the absence of the reading passage. Of course Katz et al. also show that a significant increase in correct responses occurs when passages are available to a control group. Hence it seems that Royer appears to have overgeneralized the importance of just item structure in concluding that multiple-choice reading tests do not measure passage comprehension. That is, if multiple-choice tests of reading did not tap passage comprehension and were solely a reflection of world knowledge and reasoning ability, then the subsequent addition of the passage should have had no noticeable effect on reading item correctness. Since Katz et al. clearly show a significant augmentation of item correctness when the passage was available, one must conclude that multiple-choice reading tests do measure passage comprehension and simultaneously tap other abilities, such as reasoning.

Royer's (1990) citation of Drum et al. (1981) also concerns the claimed importance of just item structure to reading comprehension item correctness. Incorrect option plausiblity was the most important predictor in Drum et al.'s

study. They classified this as an item variable. However we claim that incorrect option plausibility is more accurately classified as a text by item interaction, and is not just an item variable. That is, in order to decide whether an incorrect option is a plausible answer or not, one necessarily must scan not only the item information but the text information as well. Hence Drum et al's best predictor is one that necessarily implicates the reading of the text. This leads us to conclude that Royer's acceptance of Drum et al.'s classification scheme led him to use their results (incorrectly, we feel) to support further his hypothesis that text comprehension does not play a crucial role in multiple-choice reading tests.

But suppose Royer's (1990) critique of multiple-choice tests is assumed to be correct. Then there is little reason to expect that the 10 variables listed under hypothesis 1 (a through j above at least as they apply to the coding of the text) will be significantly related to multiple-choice reading test item difficulty. This should follow because, by (Royer's) hypothesis, multiple-choice tests are not tests of comprehension; thus variables, known to be related to comprehension difficulty (in the experimental literature), should not correlate with performance on multiple-choice reading comprehension tests. However, if Royer is incorrect, there is good reason to suppose that most if not all of the ten variables listed under hypothesis 1, at least as applied to the coding of the text, will be found to significantly correlate with reading item difficulty as obtained from multiple-choice testing.

If supporting evidence is found for hypothesis 1, there is a second implication that is important to evaluate. There are few studies that assess the simultaneous influence of many variables on comprehension (Goodman, 1982). Furthermore, many of the text materials that are evaluated in the experimental literature are not naturalistic texts but rather are artificially constructed to test the effect of one or two variables (see Hare et al, 1989, for a related argument). With the current GRE passages that are selected from naturalistic texts, it should be possible to evaluate via regression analyses whether the ten categories of variables of hypothesis 1 contribute independent information in accounting for reading comprehension item difficulty. This leads us to our second hypothesis.

Hypothesis 2. Many of the 10 categories of variables provide independent predictive information in accounting for reading item difficulty.

Corollary to hypothesis 2. Confirmation of hypothesis 2, using GRE data, implies that many of the nine categories of variables for hypothesis 1 apply to naturalistic texts as well as to the more controlled texts employed in any experimental studies of reading comprehension.

Materials and Method

The 244 reading comprehension items taken from 28 GRE forms comprise the total item sample. The total number of reading passages represented was 110. Each test form consists of four reading passages, two long (400 words or more) and two short (200 words or less). Only main idea (n=76), inference (n=87), and explicit statement (n=81) items were selected for study. Other item types such as author's tone and author's organization occur infrequently and were not scored because subanalyses of each item type were planned. We also did not sample items that use a Roman numeral type format (e.g., where different combinations of three elements comprise the list of options as in (a) only I is correct, (b) only I and II are correct, (c) I and III are correct, (d) II and III are correct, (e) none are correct). We also excluded special items that feature a capitalized NOT or LEAST in the item stem.

The data for each item difficulty measure (equated delta) was based on approximately 1,800 examinees; these examinees were randomly selected from a much larger pool of examinees who responded to each GRE test form. The equated delta value slightly adjusts the difficulty of each item across forms so that items can be meaningfully compared across forms. The adjustment stems from the fact that the sample of examinees who respond to a particular test form differ slightly in overall ability level from those responding to other test forms. The delta of each test form is adjusted so it has a mean of 13.0 and a standard deviation of 4.0.

The analysis of each item type was restricted in the following way. Each passage that was selected for a particular analysis was associated with a single exemplar of each of the three reading item types. Thus, for 76 main idea items there were 76 passages; for 87 inference items there were 87 passages; for 81 explicit statement items there were 81 passages. Sometimes the same passage was a source for several of the item types, and sometimes only one item was associated with a particular passage. For all analyses reported below, only single item types were involved. These restrictions in sampling were undertaken to avoid what statisticians call a "nesting" effect.

Most of the independent variables listed below were motivated by the literature review presented above. These, along with a few additional variables (e.g., number of rhetorical questions in the passage, type of passage subject matter, lexical coherence across text paragraphs), had been used in an earlier study of SAT reading item difficulty using a multiple-choice testing format (Freedle & Kostin, 1990).

Independent Variables for Representing Text and Item Information

Item variables

Item type

v1--Main idea
v2--Inference
v3--Explicit statement

Variables for item's stem

v4--Words in Stem: Number of words in stem (the item question)
v5--Hedges in Stem: Use of hedge (e.g., perhaps, probably) in stem
v6--Fragment Stem: Use of full question, or, sentence fragment
v7--Negative Stem: Use of simple negation
v8--Fronted Stem: Use of fronting (e.g., use of any phrases or clauses preceding the subject of the main independent clause, or use of clefts--see below under text variables for details)
v9--Reference Stem: Sum of referentials to text, stem,or options (see below for definitions under text variables)
v10--Reference Line Stem: Reference made to text lines or paragraphs

Variables for item's correct option

v11--Answer Position: Ordinal position of correct answer
v12--Words Correct: Number of words in correct option
v13--Negative Correct: Use of simple negation(s) in correct option
v14--Fronting Correct: Use of fronting(s) in correct option
v15--Reference Correct: Use of referential(s) in correct option

Variables for item's incorrect options

v16--Words Incorrects: Number of words summed over all incorrect options
v17--Negative Incorrects: Use of simple negation(s) summed over incorrect options
v18--Fronted Incorrects: Use of fronting(s) summed over incorrect options
v19--Reference Incorrects: Use of referential(s) summed over incorrect options

Text Variables

Vocabulary variable for text

v20--Vocabulary: Number of words with three or more syllables for the first 100 words of the passage (estimates vocabulary difficulty)

Concreteness/abstractness of text

v21--Concreteness: Determines whether main idea of text and its development are concerned with concrete or abstract entities.

Subject matter variables of text

v22--Physics

v23--Biology

v24--Natural science: Combined v22 and v23 into a single natural science variable

v25--Social Science: Subjects such as history, anthropology, economics, sociology, political science

v26--Humanities: Subjects such music, architecture, literary criticism, philosophy

v27--Natural science excerpt: Represents an "excerpt of natural science"

v28--About natural science: Represents a passage "about natural science"

Type of rhetorical organization

v29--Argument: Rhetorical presentation (i.e., author favors one of several points of view presented in text; occasionally other viewpoints may be only implied)

v30--List/Describe: Grimes' (1975) rhetorical organizer that interrelates a collection of elements in a text that are related in some unspecified manner; a basis of a list "... ranges from a group of attributes of the same character, event, or idea, to a group related by simultaneity to a group related by time sequence" (Meyer, 1985, p. 270). Describe relates a topic to more information about it. We felt this was sufficiently similar to list to warrant scoring them as members of the same category.

v31--Cause: Another Grimes (1975) rhetorical organizer. "Causation shows a causal relationship between ideas where one idea is the antecedent or cause and the other is a consequent or effect. The relation is often referred to as the condition, result, or purpose with one argument serving as the antecedent and the other as the consequent. The arguments are before and after in time and causally related." (Meyer, 1985, p. 271).

v32--Compare: Another Grimes (1975) rhetorical organizer. The comparison relation points out differences and similarities between two or more topics.

The two subtypes of compare used here are as follows:

v33--Compare-adversative (this relates a favored view to a less desirable opposing view) and

v34--Compare-alternative (this interrelates equally weighted alternative options or equally weighted opposing views) (Meyer, 1985, p. 273).

v35--Problem/solution: This is defined as follows: "... similar to causation in that the problem is before in time and an antecedent for the solution. However, in addition there must be some overlap in topic content between the problem and solution; that is, at least part of the solution must match one cause of the problem. The ... problem and solution ... are equally weighted and occur at the same level in the content structure" (Meyer, 1985, p. 272).

Coherence of lexical concepts over whole text

v36--Coherence (this involves judging whether opening concepts of the first sentence occur throughout the text paragraphs: 3 = maximum lexical coherence, ... 0 = no obvious lexical overlap).

Lengths of various text segments

v37--Paragraphs: Number of passage paragraphs
v38--Text words: Number of words in passage
v39--Text sentences: Number of text sentences
v40--First paragraph words: Number of words in first paragraph
v41--Longest paragraph words: Number of words in longest paragraph
v42--First paragraph sentences: Number of sentences in first paragraph
v43--Longest paragraph sentences: Number of sentences in longest paragraph
v44--Text sentence words: Average number of words per text sentence
v45--Text paragraph words: Average number of words per paragraph
v46--First paragraph sentence length: Average length of sentences in first paragraph
v47--Longest paragraph sentence length: Average length of sentences in longest paragraph

Occurence of different text "frontings": v48-v54 distinguish several types and combinations of "frontings." Some examples follow.

Use of theme-marking: In the front, the car rocked.
Fortunately, the car rocked.
Use of coordination: But, the car rocked.
Use of clefts (deferred foci): It is the case that George is short.
There are cases that defy reason.
(It and there function as dummy elements without a referent.)
Use of combinations: And, near the rear, the toy fell.
Longest run of frontings: Number of successive independent clauses that begin with fronted information (e.g., "The man laughed. Then, he frowned. And when he turned, fell." This example of three independent clauses has two successive sentences with fronted material; hence its run length is "2".).

v48--Percent fronted paragraph openings: Percentage of fronted clauses in the opening clauses across all paragraphs
v49--Frequency fronted paragraph openings: Frequency of fronted clauses in the opening clauses across all paragraphs
v50--Percent fronted text clauses
v51--Frequency fronted text clauses
v52--Frequency combinations of fronted text structures
v53--Frequency of text clefts: this is sometimes referred to as deferred foci which is one type of fronting)
v54--Longest fronted run: number of consecutively fronted text clauses

Text questions

v55--Text questions: Number of rhetorical questions in text

Text referentials

v56--Reference within text clauses: Frequency of within-clause referentials of all text clauses (e.g., "When George fell, he hurt.")
v57--Reference across text clauses: Frequency of across clause referentials (e.g., "George fell. That hurt.")
v58--Frequency special reference: Reference outside text (e.g., "One might feel sorry for George.")
v59--Reference Sums: Sum of v56, v57, v58

Text negations

v60--Text negatives: Number of simple negations in text

Special text by item interactions: the location of text information relevant to answering a particular item correctly. (Note: many item stems and/or item options specify a specific content to be searched for in the text [e.g., "according to the text, when the author said x, this means ...."]: scoring where in the text this linkage of critical stem information occurs has been designated as a text by item interaction variable.)

Text by item interactions applicable only to main idea information

v61 to v69--in general these variables specify location of main idea information at various places in the surface text.

v61--Main idea first sentence: Main idea information is in first sentence of text
v62--Main idea second sentence: Main idea information is in second sentence of text
v63--Main idea first short paragraph: Main idea information is in first short paragraph (100 words or less, excluding instances of v61 & v62)
v64--Main idea opening second paragraph: Main idea information in first sentence of second paragraph
v65--Main idea middle text: Main idea information is near middle of passage
v66--Main idea final short paragraph: Main idea information is in last short paragraph (100 words or less in paragraph, excluding instances of v41)
v67--Main idea last text sentence: Main idea information is in last sentence of text
v68--Main idea no specific location: Main idea information is not located in any specific part of the text
v69--All early main idea locations: Sum of v61, v62, and v63. Several of the analyses below used only this combined category--that is, v69=v61+v62+v63--since this was found to improve predictability of some of the criterion variables in our earlier reading study (Freedle & Kostin, 1990).

Text by item interactions applicable to inferences and explicit statement items

v70--Easily found word same sentence: Stem sends you to unique easily found word in text and relevant information is in that same sentence. (Easily found means that word stands out from text

because it is in caps or quotes or involves special letters [as in "CS103"].)

v71--Easily found word next sentence: Stem sends you to unique easily found word in text but relevant information is in next sentence.

v72--Unique word same sentence: Stem sends you to unique word in text (but it is not easily discriminated from rest of text) and relevant information is in same sentence.

v73--Unique word next sentence: Same as v72 except relevant information is in next text sentence.

v74--Unique word previous sentence: Same as v72 except relevant information is in previous text sentence.

v75--Unique word later sentence: Same as v72 except relevant information is later in same paragraph, not the next sentence.

v76--Unique word earlier sentence: Same as v72 except relevant information is much earlier in same paragraph.

v77--Unique word different paragraph: Same as v72 except relevant information is in a different paragraph.

v78--Key word multiple places: Stem suggests a particular topic but that topic is mentioned more than once.

v79--Information in first sentence: Relevant information is in first text sentence

v80--Information in second sentence: Relevant information is in second text sentence

v81--Information in first short paragraph: Relevant information is in first short paragraph of 100 words or less (but not in first two sentences).

v82--Information in opening second paragraph: Relevant information is in first sentence, second paragraph

v83--Information in last sentence: Relevant information is in last sentence of text.

v84--Information last short paragraph: Relevant information is earlier as in a last short paragraph of 100 words or less (but is not in last sentence).

v85--Information middle of text: None of v79 to v84, but is relevant information located more in middle of text.

v86--Information from two paragraphs: Relevant information must be integrated from two text paragraphs.

v87--No directive stem information: No information in stem leads you to a specific place in text (e.g., stem reads "According to the passage which of the following statements is true: ...").

v88--Words before critical information: Number of words in passage you have to read before the sentence containing the relevant information begins.
v89--Words in relevant paragraph: Number of words in paragraph in which the relevant information is located.
v90--Information middle relevant paragraph: Relevant information is in middle of a paragraph rather than the first or last sentence of that paragraph.
v91--Sum early information codes: that is, v79 + v80 + v81.

Dependent variable

v92--Item difficulty: Item equated delta (referred to as just "delta")

The dependent variable is an item's equated delta (an item's difficulty that converts percent corrects per test form to a common scale with mean of 13.0 and standard deviation of 4). See above for a more detailed description of equated delta.

In scoring items, the structure and content of item stems, correct options, and incorrect options were recorded using the 19 variables listed above (3 of these 19 being the code for item type). A related set of variables was scored for capturing the passage information but included additional variables that were unique to the text structure (see variables listed above). In all there are 39 text variables that apply to each of the three item types. Also there are 9 text by item variables for main idea items; for inference as well as explicit statement items there are 22 text by item variables.

Results and Discussion

Table 1 presents data that help to identity those variables that will be important in predicting reading item difficulty. In Table 1 we see that 42 different variables yield a significant correlation with item difficulty (equated delta). First we will use portions of Table 1 to assess the apparent adequacy of hypothesis 1 for each of the 10 categories listed under the hypothesis. We are primarily interested here in how well the text and text related variables satisfy hypothesis 1 (because of the Royer, 1990, critique of multiple-choice tests of reading comprehension); significant effects of these categories for item variables, however, will also be pointed out. Also, we do not predict that each reading item type should necessarily reveal a significant correlation with each of the 10 categories listed under hypothesis 1; we do expect, however, to find some evidence, pooled over all three reading item types, that establishes the argument that a multiple-choice format yields findings similar to those reported in the experimental literature where other response formats (such as recall) have generally been employed.

Correlates of Reading Item Difficulty as Determined by the Categories of Hypothesis 1 a. As expected, text negations (v60--text negatives) do significantly influence comprehension difficulty. Main idea items significantly correlate with text negations in the expected direction--the more text negations the harder the main idea item. However, for the item variables we see that inference items correlate significantly with the item's correct option negations (v13--negative correct)--but the negative sign obtained for correct option negations is in the opposite direction of that expected. Since, for inference items, the same negative sign occurs for the item's stem negations (v7--negative stem), it might be that a matching operation across the item's stem and its correct option is, in part, accounting for this facilitative effect of negations; this will have to be studied in more detail elsewhere. These latter unexpected findings with respect to the sign of the correlation coefficient for negatives is the only place in this study where the findings contradict the expected directional prediction based on the literature review. So, regarding negations, only the text negations relate as expected to reading difficulty while the item negations correlate significantly but in the opposite direction of that expected. Hence negations cannot be counted as either confirming nor disconfirming one of the 10 categories under hypothesis 1.

b. As predicted, the number of text referentials is significantly related to reading difficulty, here for main ideas, for variables v57, v58 and v59--reference across text clauses, frequency special reference, and reference

Table 1

Correlations of Significant Item and Text Variables with Equated Delta for Three GRE Reading Item Types

Variable		Significant Correlation of Delta with Three Reading Item Types (n=76) Main Idea[a]	(n=87) Inference	(n=81) Explicit
v7	Negative stem	--	-.20*	--
v12	Words correct	--	--	.29***
v13	Negative correct	--	-.28***	--
v14	Fronting correct	--	.25**	.19++
v15	Reference correct	--	--	.23**
v16	Words incorrects	.22*	.21**	.28***
v19	Reference incorrects	--	--	.30***
v21	Concreteness	--	-.21**	-.27**
v24	Natural science	--	--	-.26**
v26	Humanities	--	--	.23**
v27	Natural science excerpt	--	--	-.30***
v29	Argument	.21++	--	--
v34	Compare-alternative	--	--	-.22**
v36	Coherence	-.27**	--	--
v41	Longest paragraph words	.18+	--	--
v44	Text sentence words	.21++	--	--
v46	First paragraph sentence length	--	--	.18++
v47	Longest paragraph sentence length	.20++	--	--
v50	Percent fronted text clauses	.29***	--	--
v51	Frequency fronted text clauses	.25**	--	--
v52	Frequency combinations of fronted text structures	.25**	--	--
v53	Frequency of text clefts	.18+	--	--
v54	Longest fronted run	.22*	--	--
v55	Text questions	.19++	--	--
v57	Reference across text clauses	.18+	--	--
v58	Frequency special reference	.31***	--	--
v59	Reference sums	.23**	--	--
v60	Text negatives	.23**	--	--
v61	Main idea first sentence	-.20++	NA	NA
v65	Main idea middle text	.27**	NA	NA
v69	All early main idea locations	-.20++	NA	NA
v70	Easily found word same sentence	NA	-.23**	-.18++
v73	Unique word next sentence	NA	--	-.21*
v77	Unique word different paragraph	NA	--	.26**
v79	Information in first sentence	NA	-.34***	-.21*
v80	Information in second sentence	NA	--	-.21*
v82	Information in opening second paragraph	NA	-.28***	--
v83	Information in last sentence	NA	-.21**	--
v85	Information middle of text	NA	.28***	.25**

Table 1 (Continued)

v89	Words in relevant paragraph	NA	--	.20*
v90	Information middle relevant paragraph	NA	.39***	.28***
v91	Sum early information codes	NA	--	-.29***

a
A positive correlation for delta means the variable makes the items harder. *** = signif. at $p < .01$, 2-tail; ** signif. at $p < .05$, 2-tail; * $p < .06$, 2-tail; ++ $p < .05$, 1-tail; + $p < .06$, 1-tail. NA = not applicable. If a variable was not significant for the 2-tail test but appeared as one of the variables listed under hypothesis 1 where direction was predicted, we applied a 1-tail test. Also if a main idea variable was not significant at the 2-tail test and it was significant for our earlier SAT main idea data (Freedle & Kostin, 1990), we again applied a 1-tail test.

sums, respectively. Also for explicit statement items we get a significant relationship for variables v15 (reference correct) and v19 (reference incorrects) which also involve referentials. In general the more text referentials present (v57, v58, v59--reference across text clauses, frequency special reference, reference sums, respectively), the harder the comprehension process is for main ideas. The number of referentials in the item structure itself (v15--reference Correct; v19--reference incorrects) influences explicit statement item difficulty by making such items more difficult.

c. In line with our general prediction, we see that at least one of the text's rhetorical organizers (v34--compare-alternative) is significantly correlated with explicit statement item difficulty: the compare-alternative organizer (v34) makes the item easier. V29 (argument) for main idea items can also be counted as among the text's rhetorical organizers.

d. As generally predicted, the number of fronted structures in the text as measured by variables v50, v51, v52, v53 and v54 (percent fronted text clauses; frequency fronted text clauses; frequency combinations of fronted text structures; frequency of text clefts; longest fronted run) make main idea items harder. V50 and v51 (percent fronted text clauses and frequency fronted text clauses) both deal with the sum of all singly fronted types (clefts and marked topics) while v52 (frequency combinations of fronted text structures) deals with combinations of fronted types such as clefts combined with coordinations. Also, we note that for the item's correct option, v14 (fronting correct) makes the item harder for both inference and explicit statement items.

e. Vocabulary (v20) did not show a significant effect contrary to our hypothesis.

f. We see that a 1-tailed test suggests that main ideas become more difficult the longer the sentences in the text are (v44--text sentence words--and v47--longest paragraph sentence length). The text variable v46 (first paragraph sentence length) also contributes to item difficulty for explicit statement items.

g. A text's longest paragraph length (v41--longest paragraph words) shows a marginally significant 1-tail test for main idea items in the expected direction. Also, later in this report, we demonstrate that for long passages (greater than 400 words) there is a significant relationship ($p < .01$, 2-tail) of a text's longest paragraph length with main idea difficulty.

h. Number of paragraphs (v37--paragraphs) did not have a significant correlation with our three reading item types.

i. As predicted, the concreteness (v21) of the text showed a significant effect; it is significant for both inference and explicit statement item types. (Concreteness of text makes these item types easier.)

j. As predicted the following location variables are significantly correlated in the expected direction with reading difficulty: v61, v65, v69, v79, v80, v82, v83, v85, v90 and v91 (main idea first sentence, main idea middle text, all early main idea locations, Information in first sentence for explicits and inferences, information in second sentence for explicits, information in opening second paragraph for inferences, information in last sentence for inferences, information in middle of text for explicits and inferences, information in middle relevant paragraph for explicits and inferences, and sum early information codes for explicits).

Excluding negations, 7 of the 10 categories listed under hypothesis 1 generally show the expected significant relationship of the text, item, and/or text by item interaction variables with one or more of the reading item types. Thus these results taken together appear to confirm, for multiple-choice testing formats, that the variables reported as important for comprehension in the empirical literature are also important contributors affecting comprehension difficulty in a multiple-choice testing format.

Concerning only the text and/or text-related variables, there were also 7 significant categories. This particular result suggests that a multiple-choice format does not interfere with assessing passage comprehension and yields results similar to those found in the experimental literature; hence these particular text correlational results call into question some of Royer's (1990) criticisms of multiple-choice tests of reading.

We next take up each item type and explore in greater detail the variables already mentioned above plus additional variables that were not included among our set of categories for hypothesis 1. In the interest of brevity we restrict our comments to just those 27 variables that yielded correlations with significance levels of $p < .05$, 2-tail, or better.

Correlational Results for Main Idea Items. The more negations in the passage (v60--text negatives) the harder it is for examinees to determine the main idea. Main idea response options are almost always stated in affirmative terms. Thus, this result suggests that an extra step may be involved in restating relevant text segments so as to agree with the positively stated

item alternatives. This extra step introduces the possibility of error and hence makes such items more difficult.

If a passage is highly coherent (v36--coherence), it is easier to get the main idea item correct. A highly coherent text repeats one or more of the lexical concepts of the opening sentence throughout all subsequent paragraphs of the passage. Also, we see that if the main idea is located in the middle of the passage (v65--main idea middle text), it makes the main idea item harder.

All remaining significant correlations ($p < .05$, 2-tail) for main idea items deal with the use of text referentials (v58--frequency special reference--and v59--reference sums) and text frontings (v50, v51, v52--percent fronted text clauses, frequency fronted text clauses, frequency combinations of fronted text structures), all of which make main idea items more difficult. If many "special" referential pronouns are included in the text (such as the indefinite "one can see that ..." or the indefinite "we know that...."), this appears to make it more difficult to clearly identify what the main idea of the passage is. Similarly, the sum of all pronominal referentials scored (within clause and across clause pronouns, including special pronouns) also appears to make main idea items harder (of course, this could be due solely to the inclusion of special pronouns in this summed score--later in the regression analyses we shall be able to partial out the redundancy from these several scores). The various text frontings also make main idea items harder--these scores reflect qualifying phrases and so forth that occur prior to the main sentence subject. Apparently this qualifying information makes the examinee less certain about what the key idea is.

Correlational Results for Inference Items. As Table 1 also shows for a 2-tail test, there are three variables that contribute to making inference items more difficult: v16 (words incorrects), v85 (information middle of text), v90 (information middle relevant paragraph). As the number of words in the incorrect option (v16--words incorrects) increases, so does the difficulty of an inference item. Presumably longer incorrect options increase the number of different concepts that have to be compared against the relevant text information; thus, this increases the number of processing steps prior to making a final decision.

If the relevant information is in the middle of the passage (v85--information middle of text) or if it is in the middle of a particular text paragraph (v90--information middle relevant paragraph), examinees find it more difficult to find this information and thereby select the appropriate response option.

The following variables appear to make inference items easier: v13, v21, v70, v79, v82, v83--negative correct, concreteness, easily found words same sentence, information in first sentence, information in opening second paragraph, and information in last sentence. If the passage has a concrete orientation (v21--concreteness), the inference item is easier; presumably the ability to visualize a concrete set of text concepts improves the precision with which an inference can be drawn. If the inference stem sends the examinee to relevant places in the text which are easy to locate (v70--easily found word same sentence) or if the relevant information is at the beginning or end of the text (v79, v82, v83--information in first sentence, information in opening second paragraph, information in last sentence), this contributes to making the inference item easier.

The only difficult finding to explain is that v13 (negative correct) facilitates making a correct inference! Our prior work using SAT reading comprehension items (Freedle & Kostin, 1990) suggested that the presence of a negation generally makes an item more difficult. It is possible, of course, that by chance the relevant text assertions may themselves be stated in the negative; if so, it would be easier to confirm a negative text statement if an item's correct option were also stated in the negative. However, we have not checked inference item instances to verify whether this conjecture is the likely explanation.

Correlational Results for Explicit Statement Items. Eight variables contribute to making explicit statement items more difficult by a 2-tail test: v12, v15, v16, v19, v26, v77, v85, v90--words correct, reference correct, words incorrects, reference incorrects, humanities, unique word different paragraph, information middle of text, information middle relevant paragraph. As the number of words increase in the correct (v12--words correct) and/or incorrect options (v16--words incorrects), the potential number of different concepts that have to be compared against the text increases; thus the item becomes more difficult. As the number of referential expressions increases among the correct (v15--reference correct) and/or incorrect options (v19--reference incorrects), there is also an increase in the amount of cognitive operations needed to locate the appropriate referential expression, and thus an increase in item difficulty (also see Clark & Haviland, 1977, for a related finding). If the relevant information is embedded in the middle of the text (v85--information middle of text) or the middle of a particular paragraph (v90--information middle relevant paragraph), or is in a different paragraph from the expected one (v77--unique word different paragraph), this also contributes to explicit statement item difficulty. Finally, if the passage belongs to the humanities (v26), it is

perhaps more difficult to locate or correctly interpret the relevant text information.

The following variables make explicit statement items easier: v21, v24, v27, v34, v91--concreteness, natural science, natural science excerpt, compare-alternative, sum early information codes. It is somewhat surprising to find that the concreteness of the text (v21--concreteness) can also facilitate the locating of explicit information. Presumably the ability to scan the details of a concrete passage rapidly in order to locate explicit information is faster and more accurate than scanning an abstract passage. Also, if the passage is a science excerpt (v27--natural science excerpt), this facilitates explicit item difficulty; however, most science excerpts turn out to be concrete passages so the significance of v27 natural science excerpt is probably not independent of the significance of v21--concreteness. Similarly, the presence of natural science content (v24--natural science) facilitates correctness; but again, most of the natural science passages are concrete and are also primarily science excerpts.

If the passage has a top-level compare-alternative structure (v34--compare-alternative), this aids getting the item correct. Finally, if the explicit information is in an easily found passage location (e.g., at the beginning of the passage, v91--sum early information codes), the explicit item is easier.

General Comparison of Three Reading Item Types. From the above correlational results we point out a few of the similarities among the three item types. All three item types show v16 (words incorrects) as significant. Text concreteness (v21--concreteness) contributes similarly to item difficulty for explicit and inference items: concrete texts are easier for both item types than are abstract texts. All three item types show similar locational effects: locating relevant information for an item in the middle of the text tends to make it more difficult--see v65 (main idea middle text) and v85 (information middle of text) for inference and explicit statement items; also see v90. Also, for all three item types, when the information occurs in the first sentence of the passage the items are easier--see v61 (main idea first sentence) and v79 (information in first sentence) for inference and explicit statement items. These locational effects are not necessarily unexpected because Kintsch's (1974) earlier theory suggests that early text information (which generally is easier to access and/or remember) is often the important main idea information as well, whereas less important information often tends to occur in the middle of a passage and thus is more difficult to access as readily.

However, in spite of the noted similarities above, cognitively it seems self-evident that main idea items should generally not be analyzed with other reading item types, especially explicit statement items (e.g., "According to the passage, x means the following ...."). That is, examining an entire text for its over-arching theme cannot be equivalent in all of its cognitive processing steps to confirming or disconfirming a particular statement in the passage (this being an explicit statement item).

In support of this assertion we can note the following differences in item features. In our sample of items, main Idea item stems never used a negation, whereas inference and explicit items showed a moderate use of negation in the stem. Main idea items virtually never employed "fronted" structures for the stems, but inference and explicit items showed a strong use of fronted structures for the item stem. Main idea stems never sent the examinee to a unique word in the text or to a specific topic or phrase in the text. However, inference and explicit item stems often mentioned a particular word or phrase to be searched for in the text. Also, main idea and inference items showed virtually no use of fronts for correct and incorrect options, whereas explicit items showed a moderate use of frontings for correct and incorrect options.

Contrasts and Similarities Among Three Reading Item Types. The above observations strongly suggest that the three item types typically differ in several structural features; nonetheless, we have also noted that the search process for locating the correct option sometimes yields convergent results among the three item types. Earlier studies (Drum et al, 1981; Embretson & Weltzel, 1987) unfortunately did not analyse separately each reading item type, probably as a consequence of their relatively small item sample sizes. Hence we believe that this current result may be the first effort that illustrates some similarities and differences among these three reading item types.

Not all of the correlations in Table 1 provide independent information concerning item difficulty. By conducting several stepwise regression analyses, we will be able to determine the following:

(a) what is the overall predictability of GRE reading comprehension item difficulty; that is, how much of the variance of the difficulty index (equated delta) can be accounted for?

(b) how many of the 10 category variables (listed under hypothesis 1) provide independent information for each of the three reading item types--this provides us with a clear test of hypothesis 2 (and its corollary).

Stepwise regression results for each of three reading item types. The next set of results examines the outcome of stepwise regression analyses for each of the three reading item types.

Criteria for admitting variables into the stepwise regressions. For all stepwise regressions, the following criteria were used for admitting variables into the regression. All relevant variables were available for possible selection. Each new variable that was admitted into the solution had to yield a significant individual F value, and, the new F values for all previously admitted variables had to be significant. If the next variable admitted showed a nonsignificant F, the previous solution was considered the final one.

Regarding the critical ratio, experts seem to differ. Some recommend, for example, for a 90 item sample, that no more than nine variables be extracted, providing each new variable yields a significant F value and all previous variables still retain a significant F value. Other experts (Cohen & Cohen, 1983) suggest that no more than three variables be extracted from a sample of 90 items. Yet others suggest that it is not the ratio of items to number of extracted variables that is so critical; of more importance is the

difference between the number of items and the number of predictor variables (C. Lewis, 1991, personal communication).

Because many of the variables selected for analysis were already reported in the literature as significantly related to reading comprehension (e.g., the 10 categories listed under hypothesis 1), we feel the less restrictive ratio is more appropriate, especially if we wish to provide a fair test of hypothesis 2 (which assesses whether many categories of variables simultaneously provide independent variance regarding item difficulty). However, for completeness, we also will indicate for each analysis below which variables would have been deleted had the more restrictive ratio (1 out of 30) been used. In addition (see notes 4, 5, 6) we also report how the regression results would be altered had only the variables that were significantly correlated with delta been used in the stepwise procedure (this being one way to restrict the number of predictor variables).

Overall Predictability of Item Difficulty: Evaluation of Hypothesis 2

Stepwise regression analysis of all main idea items. As Table 2 demonstrates, three significant variables (v58, v44, v46--frequency special reference, text sentence words, first paragraph sentence length) account for 20% of the variance of main idea difficulty. (The more restricted selection ratio of 1 out of 30 actually results in the same three predictors being extracted as reported in Table 2.) Thus, it appears, without additional analyses, that only 20% of the main idea item difficulty index (equated delta) can be accounted for; while significant ($p < .01$) this does not compare favorably with the 58% of the main idea variance accounted for when SAT reading comprehension items were examined (see Freedle & Kostin, 1991). Momentarily we conduct subanalyses that attempt to explore the source of these differences across the SAT and GRE main idea data.

Regarding hypothesis 2, the three extracted variables for the 76 main idea items suggest that at least two categories listed under hypothesis 1 contribute independent information concerning item difficulty: v58 (frequency special reference) representing the referential category and v44 (text sentence words) plus v46 (first paragraph sentence length), both representing the category of sentence length. (We shall have to await the regression analyses of the remaining two reading item types before we can fully evaluate the degree to which hypothesis 2 holds.)

Another issue should be pursued before we examine the other two reading item types. Because of the somewhat low predictability for all GRE main idea items, we now explore whether part of the low predictability might be due to

Table 2

Stepwise Regression Analysis for Predicting 76 GRE Main Idea Item Difficulty (Equated Delta)

Variable [a,b]	F value of each Predictor	Percent Variance	Source
All passages: (n=76 items)			
v58 Frequency special reference	8.8	9%	text
v44 Text sentence words	9.2	6%	text
v46 First paragraph sentence length	3.9	4%	text
Long passages: (n=38 items)			
v41 Longest paragraph words	15.5	18%	text
v65 Main idea middle text	5.4	9%	text by item
v43 Longest paragraph sentences	8.8	8%	text
v55 Text questions	4.7	9%	text
v51 Frequency fronted text clauses	4.2	6%	text
Short passages: (n=38 items)			
v36 Coherence	9.8	23%	text
v52 Frequency combinations of fronted text structures	6.0	11%	text
v16 Words incorrects	4.0	7%	item

a

For all main idea items the overall $F(3,72) = 5.9$, $p < .01$. Multiple R for main idea:= .444, R Squared = .20. Main idea items for long passages only: overall $F(5,32)=6.9$, $p < .01$, Multiple R= .72, R squared = .52. Main idea items for short passages only: overall $F(3,34)=7.7$, $p < .01$, Multiple R= .64, R squared = .41.

b

All F values for the individual variables are significant at $p < .05$ or beyond. These values are taken from the best stepwise regression solution.

the fact that we combined both short and long passages in the same analysis, possibly masking predictability by combining potentially different search strategies as a function of passage length. GRE passages are very clearly divided into short and long passages. If a passage is less than 200 words, it is classified as short; if longer than 400 words, it is long. It is clear that a substantial gap in length occurs between long and short GRE passages such that no passage is between 200 and 400 words in length.

The reader is reminded that Kieras' (1985) original study of identifying where the main idea of a passage is was based on single paragraphs, not multiple paragraphs. It occurred to us that perhaps by combining multiple paragraph passages with single paragraphs (these latter are virtually all defined by having fewer than 200 words in our GRE sample) this could be the source of low predictability for the full sample of main idea items.

Thus a separate regression for main ideas was conducted on just the short and just the long passages. Table 2 also shows the results of these additional stepwise regressions. The amount of delta variance accounted for appears to be substantially improved. Fifty-two percent of the variance of items associated with the long passages can now be accounted for as opposed to 20% for the full (long and short) main idea sample. For the short passages, 41% of the variance for main idea deltas now can be accounted for. But, interestingly, it is the long passages that contain a significant Kieras-type result (i.e., middle-of-passage difficulty effect, v65 [main idea middle text]), not the short passages. Perhaps the very brevity of the short passages diminishes the impact of a middle-of-passage effect. Subdividing by passage length clearly seems to have improved main idea predictability although the small sample sizes means that we cannot place too much confidence in this result.

We should note that using the more restricted selection ratio of 1 out of 30 would alter our analyses of the long and short passages for main idea items. One variable would be allowed for each subanalysis. For long passages this would result in 18% variance being accounted for by v41 (longest paragraph words--this is the same variable that emerged as one of the significant independent predictors of our SAT main idea sample, see Freedle & Kostin, 1991). For the short passages, reporting just the best predictor variable (v36) would account for 23% of the variance. Again this variable (v36--coherence) was one of the significant independent predictors of our SAT main idea sample (Freedle & Kostin, 1990).

A comparison of significant differences for the correlational results of long and short main idea subsamples (see Appendix) indicates that there is

further empirical evidence justifying a separate analysis by length of passage. For example, there were the following significant differences between long and short passages in the correlations (McNemar, 1956, $p < .05$, 2-tail for all comparisons) of the following variables with delta: v36 (coherence), v37 (paragraphs), v12 (words correct), v16 (words incorrects), v41 (longest paragraph words), and v45 (text paragraphs words). Of course it is not surprising to find v37, v41 and v45 (paragraphs, longest paragraph words, and text paragraphs words) showing a significant difference as a function of passage length; what is more interesting is the difference for v36, v12 and v16 (coherence, words correct, words incorrects). Coherence (v36) affects short passage item difficulty more than long passage item difficulty. (Mean coherence for long and short passages incidentally does not differ significantly: for short passages mean coherence = 1.74 [SD=1.13]; for long passages = 1.54 [SD=1.13].) We do not have a clear explanation for why coherence significantly affects main idea difficulty for short passages but not for long. Also, the difference between long and short passages concerning the effect of variables v12 (words correct) and v16 (words incorrects), is not obvious. For short passages, as the options become longer (i.e., use more words) the item becomes more difficult. And for long passages, as the options become longer the item becomes easier. We do not have a clear explanation for this.

Nevertheless, these results, especially for the long and short passages, must be counted as exploratory inasmuch as we have too few main idea items and too many predictor variables to guarantee a stable result. Later in this report, we show that in spite of these limitations, there is evidence that main idea reading items from the GRE overlap considerably with SAT main idea reading items regarding which variables correlate significantly with delta; however, close inspection of the findings suggests that it is the longer GRE passages that primarily account for this similarity across data sets. Such replication does add weight however to the current set of findings.[4]

Stepwise Regression Results for Predicting Inference Item Difficulty

In Table 3 we see that seven variables yield independent information concerning inference item difficulty. Jointly they account for 49% of the item difficulty variance. The following contribute to evaluating the overall predictability of inference item difficulty with respect to the categories of hypothesis 1: Concreteness (v21) and location (v90, v79, and v82--information middle relevant paragraph, information in first sentence, Information in opening second paragraph). So inferences contribute two additional significant categories favoring hypothesis 2. Incidentally, using

Table 3

Stepwise Regression Analysis for Predicting GRE Inference Item Difficulty

Variable [a,b]		F value of each predictor	Percent Variance	Source
All passages: (n=87 items)				
v90	Information middle relevant paragraph	22.0	15%	text by item
v79	Information in first sentence	15.2	13%	text by item
v82	Information in opening second paragraph	9.9	5%	text by item
v7	Negative stem [c]	6.2	4%	item
v16	Words incorrects	7.2	4%	item
v21	Concreteness	5.6	3%	text
v86	Information from two paragraphs	4.6	3%	text by item

a
Overall $F(7,79) = 10.8$, $p < .01$.
Multiple R = .70, R Squared = .49.

b
All F values for individual variables are significant at $p < .05$ or beyond.

c
This result for v70 is the only place in our regression analyses where the direction of the relationship is opposite to the expected.

the more restricted selection ratio of 1 out of 30 would mean that only the first three variables reported in Table 3 would have been counted as significant, accounting for 33% of the variance.[5]

Stepwise Regression Results for Predicting Explicit Statement Item Difficulty

Table 4 shows that seven independent predictors account for 41% of the item difficulty variance. Of those, the following variables relate to hypothesis 2: v34 (compare-alternative) the rhetorical organizer, v90 (information middle relevant paragraph--this is a location variable), v14 (fronting correct) and v15 (reference correct). Of these four categories two are new to our list of those that favor hypothesis 2 (rhetorical organizer and fronting). Finally, using the more restricted selection ratio of 1 out of 30, only the first three variables in Table 4 would have been counted as significant, accounting for 24% of the variance.[6]

Conclusion regarding evidence favoring hypothesis 2. Hypothesis 2 deals with the overall predictability of the three item types with respect to each of the 10 categories listed under hypothesis 1. Six were found to provide independent information when pooling across the three reading item types: referentials, sentence length, concreteness, rhetorical organization, frontings and location. Therefore, we conclude that there is moderate evidence favoring hypothesis 2. Thus many of the variables reported in the experimental literature as individually implicated in influencing comprehension difficulty are also found here to contribute jointly to determining comprehension difficulty distributed over several different reading item types. The corollary of hypothesis 2 (that naturalistic texts will also exhibit a similar set of jointly significant categories) also then receives moderate support.

A Comparative Analysis Using SAT Main Idea Reading Items: Further Evaluations of Hypotheses 1 and 2.

An examination of the SAT reading data (Freedle & Kostin, 1991) suggests considerable agreement with hypotheses 1 and 2 that we have just examined for the GRE data. In particular, the following categories supported hypothesis 1 for the SAT data (see column 1 of Table 5): for SAT main idea items the correlations showed negations, referentials, rhetorical organizers, frontings, paragraph length, location, and abstractness were significant. For all three SAT reading item types combined, the following categories were correlationally significant for hypothesis 1: negations, referentials, rhetorical organizers, paragraph length, location, and abstractness. All but one (negations) of

Table 4

Stepwise Regression Analysis for Predicting GRE Explicit Statements
Item Difficulty

Variable	F value	Percent Variance	Source
[a,b]			
All explicits: (n=81 items)			
v27 Natural science excerpt	11.2	9%	text
v77 Unique word different paragraph	10.5	8%	text by item
v34 Compare-alternative	9.8	7%	text
v90 Information middle relevant paragraph	9.1	5%	text by item
v14 Fronting correct	3.9	5%	item
v24 Natural science	5.8	3%	text
v15 Reference correct	5.1	4%	item

[a]
Overall F = F(7,73)= 7.4, $p < .01$.
Multiple R = .64, R Squared = .41.

[b]
The F values for individual variables are all significant at $p < .05$ or beyond.

these categories were also found for the GRE. Hence there is a significant overlap in the results of both the SAT and GRE data sets.

Regarding hypothesis 2, the SAT data supported a maximum of five independently contributing categories as jointly influencing the regression analysis predicting main idea item difficulty. This should be compared with the two categories (length and reference) which jointly supported the total main idea GRE sample. [If we scan which categories were significant for either the long or short passages for the GRE main idea regression analyses, then the number of independent categories is three (sentence length, location, frontings) rather than two.] Overall, it appears that hypothesis 2 is modestly supported by both the GRE and SAT main idea data.

Detailed Comparisons of SAT Reading Items and GRE Reading Items for Main Idea Item Difficulty: Correlational Comparisons

The main purpose of this section is to see which significant variables reported by Freedle & Kostin (1991) for the SAT main idea reading items replicate for the GRE main idea reading items. (Main idea items were scored using an identical set of variables for both the SAT and GRE reading items.)

We see in Table 5 that 22 SAT main idea variables were found to be significant ($p < .05$, 2-tail or better). If we look at the column to the extreme right (the total GRE main idea sample), we see that 10 of these individual variables were replicated as significant for the GRE total main idea sample: v29, v36, v41, v53, v55, v57, v59, v60, v65, v69 (argument, coherence, longest paragraph words, frequency of text clefts, text questions, reference across text clauses, reference sums, text negatives, main idea middle text, all early main idea locations). Of these the following 6 were predicted by hypothesis 1 to be significant: v41 (longest paragraph words), for paragraph length; v29 (argument) which implicates the Grimes rhetorical organizers, especially compare-adversative; v53 (frontings, here, text clefts); v57 (reference across text clauses) and v59 (reference sums) for the referential category; v60 (text negatives) for the negation category; v65 (main idea middle text) and v69 (all early main idea locations) for the location category.

Another way to show the similarities between SAT and GRE main ideas is to correlate the correlations themselves. The first two columns correlate .66 ($p < .002$, 2-tail), while the first and third columns correlate only .43 ($p < .05$, 2-tail), showing that the relationships associated with the longer GRE passages correlate better with the SAT than do the short GRE passages.

Table 5

Significant SAT Correlations of Item Difficulty with the Set of Variables as Compared with GRE Correlations: Main Idea Items Only

Var.	Main Idea Significant Variables for SAT (n=110)	GRE Main Idea Data: Main Idea Long Passages for GRE (n=38)	GRE Main Idea Data: Main Idea Short Passages for GRE (n=38)	GRE Main Idea Data: All Main Ideas (long & short) for GRE (n=76)
[a]	[b]			
v11	.19**	-.07	-.06	-.07
v13	.26***	-.11	-.08	-.10
v21	-.54***	-.02	-.24	-.11
v24	-.32***	-.07	-.16	-.10
v27	-.37***	-.07	-.05	-.05
v28	.20**	--	.06	.03
v29	.39***	.31*	.11	.21++
v30	-.28***	.05	-.17	-.03
v33	.38***	.32**	-.03	.12
v36	-.24***	-.04	-.48***	-.27**
v40	.22***	.10	-.22	-.07
v41	.27***	.43***	-.24	.18+
v42	.20**	.12	-.22	-.08
v43	.28***	.12	-.25	-.04
v45	.19**	.35**	-.25	.04
v53	.20**	.15	.14	.18+
v55	.28***	.22	.09	.19++
v57	.25***	.10	.26+	.18+
v59	.24***	.22	.22	.23**
v60	.35***	.32**	.05	.23**
v65	.22**	.39**	.02	.27**
v69	-.25***	-.28++	-.08	-.20**

[a]
Key for interpreting significant variables:
v11 (Answer position)
v13 (Negative, correct)
v21 (Concreteness)
v24 (Natural science)
v27 (Natural science excerpt)
v28 (About natural science)

v29 (Argument)
v30 (List/Describe)
v33 (Compare-adversative)
v36 (Coherence)
v40 (First paragraph words)
v41 (Longest paragraph words)
v42 (First paragraph sentences)
v43 (Longest paragraph sentences)
v45 (Text paragraphs, words)
v53 (Frequency of text clefts)
v55 (Text questions)
v57 (Reference across text clauses)
v59 (Reference sums)
v60 (Text negatives)
v65 (Main idea middle text)
v69 (All early main idea locations)

b

*** Correlation significant, $p < .01$, 2-tail
** Correlation significant, $p < .05$, 2-tail
* Correlation marginally significant, $p < .06$, 2-tail
++ Correlation signif., $p < .05$, 1-tail (justified if we use earlier findings of SAT to predict sign of GRE correlation).
+ Correlation is marginally significant, $p < .06$, 1-tailed.

Interestingly, the correlation of the SAT with all GRE main ideas is also significant r=.65 ($p < .002$, 2-tail), but this fails to capture the fact that the magnitudes of the largest correlations in the last column are somewhat less impressive than those found in the second column.

There is yet a third way to try to determine the similarity of the two data sets presented in Table 5. By comparing just the algebraic sign of the correlations for long passage GRE (column 2) with SAT (column 1), we find that 18 of the 21 values have the same algebraic sign; by a sign test this is significant ($p = .002$, 2-tail). However, for the short passage GRE (column 3) compared with SAT, only 14 out of 22 are in same direction; this is not significant ($p > .20$). For the full main idea GRE sample compared with the SAT main idea sample, 17 out of 22 are in the same direction, which again is significant ($p = .016$, 2-tail, sign test), this significance is due primarily to the contribution of the long GRE passages.

Conclusion

In this study we have been primarily interested in determining how well reading item difficulty can be accounted for by a set of predictors that reflect the contribution of the text structure, the item structure and the joint effect of both the text and items. We found that a substantial amount of the variance can be accounted for by a relatively small set of predictors; the range of variance accounted for varied from 20% up to 52% depending upon the particular analysis undertaken. To our knowledge this is one of the few studies to examine the predictability of a relatively large sample of multiple-choice reading items (n=244) using a wide selection of predictor variables.

Within this broader concern we have also focused upon a small set of hypotheses so as to more clearly come to terms with a number of claims that have been made in the scholarly literature concerning reading comprehension and the adequacy of reading comprehension tests per se. In particular, Goodman (1982) has complained that many of the experimental studies of comprehension have focused on just one or two variables at a time; he questions whether these separate studies taken together necessarily build up our understanding of how full comprehension of text takes place. A related concern has questioned whether the often highly artificial texts studied in the experimental literature will necessarily clarify how more naturalistic texts are comprehended. Finally, Royer (1990) and Katz et al (1990) have questioned whether multiple-choice reading tests can be considered appropriate tests of passage comprehension in light of the fact that item content alone (in the absence of the reading passage) can be demonstrated to lead to correct answers well above chance levels of guessing.

In response to these several concerns, the prediction of reading item difficulty has been framed around two hypotheses meant to put into clearer perspective the viability of multiple-choice reading comprehension tests, here exemplified by the GRE reading passages and their associated items. Since many of the scored variables deal with text content similar to those of concern in the experimental literature and since the GRE reading passages are adaptations of prose from naturalistic sources (book passages, magazines, etc.), we reasoned that the successful prediction of reading item difficulty would allow us to draw several important conclusions.

The first hypothesis asserts that multiple-choice items will be sensitive to a similar set of variables that have been found to be important in studying comprehension processes in the experimental literature. The evidence generally was interpreted to support most of the categories detailed

under hypothesis 1 for the text and text related variables. This was interpreted to mean that multiple-choice response formats yield similar results to those found in the more controlled experimental studies. Hence we feel Royer's (1990) statement that multiple-choice tests do not measure passage comprehension can be called into question.

A second hypothesis asserts that many of the significant variables will be found to jointly influence reading item difficulty. By pooling the stepwise regression results across the three reading item types, we concluded that there was considerable evidence that many of the different categories of variables studied do jointly account for reading item difficulty. This result was further interpreted as a response to Goodman's (1982) concern that since many of the experimental studies involve just one or two variables at a time, this may not be sufficient to guarantee that these variables when jointly studied will provide any cumulative new information about reading comprehension difficulty. Our results appear to suggest that in fact many of the different categories of variables do provide independent predictive information; hence the few variables studied across disparate studies in fact jointly combine so as to increase our understanding of what influences comprehension difficulty. A related set of analyses using a large number (n=285) of SAT reading items (Freedle & Kostin, 1991) further confirmed the viability of this demonstration.

The fact that the GRE passages were selected from naturalistically occurring passages was further interpreted as evidence that the predictive success of many of the variables found here to predict the difficulty of items associated with these more naturalistic passages are similar to those variables found to predict the difficulty of artificially constructed passages (as is true of many passages in the experimental literature). Thus there do not seem to be any large differences between studies using naturalistic versus artificially constructed passages in terms of their adequacy to study the factors that influence comprehension difficulty. A similar result was obtained with our analyses of SAT data (see Freedle & Kostin, 1991); because the SAT passages are also developed from naturalistically occurring prose passages, this again indicates that the distinction between artificially constructed passages and naturalistic ones is not that great in terms of assessing factors that influence reading comprehension.

In short, we find considerable evidence that multiple-choice tests of reading comprehension yield results that are quite consistent with those obtained from controlled experimental studies dealing with language comprehension. More importantly, because of the relatively large size of our data base, the results also provide evidence that many variables affecting

comprehension can be shown to contribute independent predictive information in determining reading item difficulty. A significant amount of the item difficulty variance has been accounted for by a relatively small number of variables for each of three reading item types. Finally, we find that the current results demonstrate considerable consistency across both the SAT and GRE data sets.

Future work should focus on augmenting, if possible, the amount of variance that can be accounted for for each reading item type. We believe the current work represents a significant demonstration of how to conceptualize the nature of the prediction of reading item difficulty.

Notes

1

While the Drum et al. (1981) study was innovative in analyzing the multiple-choice testing process into its constituent parts (i.e., determining the relative contributions of the item's stem, the item's correct and incorrect options, and the text variables to item difficulty), some of the study's analyses appeared to be flawed. Ten predictor variables were extracted from very small reading item samples (varying between 20 and 36 items) taken from seven children's reading tests. At most one or two predictors instead of 10 should have been extracted from such small samples (see Cohen & Cohen, 1983); hence 70% of the item difficulty variance is probably too large an estimate of the variance actually accounted for.

2

The possibility of a curvilinear relationship between v92 (item difficulty) and each of the predictor variables was examined; there was little evidence to suggest the existence of any strong curvilinear relationships in the current set of data.

3

One might postulate that the reason that items dealing with middle text information are generally more difficult than those with early text information is that there is more material to be remembered and processed that might be relevant to some particular test item. However, such a straightforward explanation would not account for the fact that test items that deal with information found in the last text sentence are often of only moderate difficulty; that is, if having to cover more text material is the source of locational difficulty, then items dealing with the final text sentence should be the most difficult items of all. But they are not the most difficult; in fact as we see for inferences (in Table 1) relevant information in the last sentence actually makes such items easier. Kintsch and van Dijk (1978) provide a memory mechanism that does account for the observed facts: it says that the reader actively processes a limited number (about four) clauses at a time; these include the most recent clause along with clauses that were judged to be of importance. Thus since the final sentence would be the most recent clause that the reader encounters upon finishing his or her reading, it would be one that is in active memory and so should a reading item deal with such information it should be relatively easy to get such an item correct.

4

Another stepwise reression was run for the main idea items using only those variables that significantly correlated with main idea item difficulty (delta). For the total main idea items v58 (frequency special reference)

v44 (text sentence words) represented the final solution (these are identical to the first two variables in Table 2 that emerged when all predictor variables were allowed to enter into the solution). V65 (frequency special reference) and v44 (text sentence words) account for 15% of the variance with $F(2,73) = 6.6$, $p < .01$. Clearly this result is similar to the 20% accounted for in Table 2. For long passages, when we allowed only significantly correlated variables into the final solution we can account for 28% of the variance, $F(2,35) = 6.7$, $p < .01$; again the same two predictor variables v41 (longest paragraph words) and v65 (main idea middle text) as reported in Table 2 were extracted. For the short passages, the best predictors of the main idea items were identical as those reported in Table 2 [v36 (coherence), v52 (frequency combinations of fronted text structures), v16 (words, incorrects]; the F value was identical. Clearly, this alternative method for restricting the total number of variables that are allowed into the regression solution yields a result very similar to the ones reported in Table 2 wherein the full set of predictor variables were allowed. Also because of the similarity of the final solutions one can deduce that there are few if any suppressor variables present in our data.

5

Another stepwise regression was run for the inference items using only those variables that significantly correlated with inference item difficulty (delta). The best seven variables were v90 (information middle relevant paragraph), v79 (information in first sentence), v82 (information in opening second paragraph), v7 (negative stem), v16 (words, incorrects), v21 (concreteness) and v13 (negative correct) which together account for 49% of the variance, $F(7,79) = 10.7$, $p < .01$. This agrees very well with the variables that were found in Table 3 wherein all predictor variables were allowed into the regression. This again shows that by restricting the allowed variables to just those that yield significant correlations with the dependent variable (delta) that the results are essentially identical to the earlier method.

6

Another stepwise regression was run for the explicit statement items using only those variables that significantly correlated with explicit statement item difficulty (delta). Here the final solution was identical to that reported in Table 4. Again this shows that one way to restrict the number of variables initially allowed in the regression solution is to sample just the significantly correlated variables. All three reading item types showed this close agreement between the the two approaches to stepwise regressions.

References

Abelson, R. P., & Black, J. B. (1986). Introduction. In J. A. Galambos, R. P. Abelson, & J. B. Black (Eds.), Knowledge structures. Hillsdale, NJ: Erlbaum.

Abrahamsen, E., & Shelton, K. (1989). Reading comprehension in adolescents with learning disabilities: semantic and syntactic effects. Journal of Learning Disabilities, 22, 569-572.

Bever, T. G., & Townsend, D. (1979). Perceptual mechanisms and formal properties of main and subordinate clauses. In W. Cooper & E. Walker (Eds.), Sentence processing. Hillsdale, NJ: Earlbaum.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82, 45-73.

Clark, H. H., & Haviland, S. (1977). Comprehension and the given-new contract. In R. Freedle (Ed.), Discourse production and comprehension. Norwood, NJ: Ablex.

Cohen, J., & Cohen, P. (1983). Applied multiple regression--correlational analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-514.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Ervin-Tripp, E. (1964). An analysis of the interaction of language, topic and listener. American Anthropologist, 66, 86-102.

Freedle, R. (1990). Evidence for the construct validity of the SAT reading comprehension items. Manuscript submitted for publication.

Freedle, R., & Duran, R. (1979). Sociolinguistic approaches to dialogue with suggested applications to cognitive science. In R. Freedle (Ed.), New directions in discourse processing. Norwood, NJ: Ablex.

Freedle, R., Fine, J., & Fellbaum, C. (August, 1981). Predictors of good and bad essays. Paper presented at the annual Georgetown University Roundtable on languages and linguistics. Washington, DC.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages using three item types: main idea, inference and explicit statement items. Princeton, NJ: ETS Research Report RR-91-29.

Goodman, K. (1982). Language and literacy: The selected writings of Kenneth S. Goodman (Vols 1 & 2). F. Gollasch (Ed.) Boston: Routledge & Kegan Paul

Graves, M. (1986). Vocabulary learning and instruction. In E. Rothkopf (Ed.), Review of research in education, Vol. 13. Washington, DC: American Educational Research Association.

Grimes, J. (1975). The thread of discourse. The Hague: Mouton.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. Journal of Educational Psychology, 79, 220-227.

Hare, V., Rabinowitz, M., & Schieble, K. (1989). Text effects on main idea comprehension. Reading research quarterly, 24, 72-88.

Hites, R. W. (1950). The relation of readability and format to retention in communication. Unpublished doctoral dissertation. Ohio State University.

Hymes, D. (1962). The ethnography of speaking. In T. Gladwin & W. Sturtevant (Eds.), Anthropology and human behavior. Washington, DC: Anthropological Society of Washington.

Johnston, P. (1983). Reading comprehension assessment: a cognitive basis. Newark, NJ: The International Reading Assoc.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. Reading Research Quarterly, 19, 219-239.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Boston: Allyn and Bacon.

Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. Psychological Science, 1, 122-127.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. In B. Britton & J. Black (Eds.), Understanding expository text. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Klare, G. (1974-1975). Assessing readability. Research Research Quarterly, 10, 62-102.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111-151.

McNemar, Q. (1955). Psychological statistics, 2nd edition. NY: Wiley.

Meyer, B. (1975). The organization of prose and its effects on memory. Amsterdam: North-Holland.

Meyer, B. (1985). Prose analysis: purposes, procedures and problems. In B. Britton & J. Black (Eds.), Understanding expository text. Hillsdale, NJ: Erlbaum.

Meyer, B., & Freedle, R. (1984). The effects of different discourse types on recall. American Educational Research Journal, 21, 121-143.

Paivio, A. (1971). Imagery and verbal processes. NY: Holt, Rinehart & Winston.

Royer, J. (1990). The sentence verification technique: a new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), Cognitive assessment of language and math outcomes. Norwood, NJ: Ablex.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning. Hillsdale, NJ: Erlbaum.

Tuinman, J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

Appendix

The Correlation between GRE Item Difficulty (delta) and Each Predictor Variable for Each of Three Reading Item Types

(only 2-tail tests reported)

Table 6

Correlation between GRE Item Difficulty
and Each Predictor Variable

Variable	All Inferences (n=87)	All Explicits (n=81)	All Main Ideas (n=76)	Long Passages Main Ideas (n=38)	Short Passages Main Ideas (n=38)
v4	-.19	.04	.10	.00	.19
v5	-.15	-.04	--	--	--
v6	-.06	-.19	-.01	.00	.02
v7	-.20*	-.06	--	--	--
v8	.02	-.15	.00	--	.04
v9	.06	.15	-.02	-.06	-.02
v10	-.06	.14	--	--	--
v11	.16	.09	-.07	-.07	-.06
v12	.11	.29***	.06	-.18	.26
v13	-.28***	.03	-.10	-.11	-.08
v14	.25**	.19	.16	.04	.27
v15	-.11	.23**	-.04	-.16	.06
v16	.21**	.28***	.22*	-.04	.41**
v17	-.10	.01	-.02	.12	-.15
v18	.06	.16	.12	.04	.28
v19	-.06	.30***	.09	-.10	.27
v20	-.14	.02	.09	.11	.06
v21	-.21**	-.27**	-.11	-.02	-.24
v22	-.10	-.05	-.01	-.05	.00
v23	.15	-.25**	-.10	-.02	-.19
v24	.04	-.26**	-.10	-.07	-.16
v25	-.14	.07	.07	.01	.14
v26	.09	.23**	.05	.07	.05
v27	.10	-.30***	-.05	-.07	-.05
v28	.04	-.06	.03	--	.06
v29	-.14	.13	.21	.31*	.11
v30	-.03	.16	-.03	.05	-.17
v31	.01	.06	.05	-.07	.15
v32	-.06	-.11	.04	.11	.03
v33	.03	.10	.12	.32**	-.03
v34	-.10	-.22**	-.05	-.19	.06
v35	.10	-.18**	-.07	-.09	-.05

Table 6 (Continued)

v36	-.12	.07	-.27**	-.04	-.48***
v37	-.09	-.07	.06	-.27	.25
v38	-.04	-.03	.13	.17	.13
v39	-.02	-.09	.06	-.21	-.14
v40	.11	.02	-.07	.10	-.22
v41	.03	.04	.18	.43***	-.24
v42	.13	-.06	-.08	.12	-.22
v43	.11	-.07	-.04	.12	-.25
v44	-.10	.12	.21	.27	.17
v45	.07	.05	.04	.35**	-.25
v46	-.04	.18	-.06	-.17	.02
v47	-.11	.14	.20	.40**	.05
v48	.02	.07	.10	.10	.08
v49	-.10	.00	.14	.05	.17
v50	-.04	-.01	.29***	.40**	.20
v51	-.09	-.08	.25**	.38**	.16
v52	-.06	-.10	.25**	.22	.25
v53	.05	-.06	.18	.15	.14
v54	-.05	-.16	.22*	.24	.12
v55	.04	.14	.19	.22	.09
v56	-.03	.12	.06	.07	-.18
v57	.04	-.06	.18	.10	.26
v58	.11	.02	.31***	.30	.30
v59	.05	.03	.23**	.22	.22
v60	.03	.11	.23**	.32**	.05
v61	NA	NA	-.20	-.26	-.11
v62	NA	NA	-.16	-.23	-.08
v63	NA	NA	-.03	-.10	.05
v64	NA	NA	-.15	-.23	-.11
v65	NA	NA	.27**	.39**	.02
v66	NA	NA	.00	-.10	.10
v67	NA	NA	-.01	.14	-.10
v68	NA	NA	.04	.11	.01
v69	NA	NA	-.20	-.28	-.08
v70	-.23**	-.18	NA	NA	NA
v71	.11	-.12	NA	NA	NA
v72	.08	-.12	NA	NA	NA
v73	.11	-.21*	NA	NA	NA
v74	-.11	-.11	NA	NA	NA
v75	-.03	-.08	NA	NA	NA
v76	.10	--	NA	NA	NA

Table 6 (Continued)

v77	.14	.26**	NA	NA	NA
v78	-.04	.17	NA	NA	NA
v79	-.34***	-.21*	NA	NA	NA
v80	.06	-.21*	NA	NA	NA
v81	.14	-.16	NA	NA	NA
v82	-.28***	-.04	NA	NA	NA
v83	-.21**	-.11	NA	NA	NA
v84	.05	.10	NA	NA	NA
v85	.28***	.25**	NA	NA	NA
v86	.16	.02	NA	NA	NA
v87	.07	.18	NA	NA	NA
v88	.12	.11	NA	NA	NA
v89	.03	.20*	NA	NA	NA
v90	.39***	.28***	NA	NA	NA
v91	-.16	-.29***	NA	NA	NA

*** significant, p <.01, 2-tailed
** significant, p <.05, 2-tailed
* marginally significant, p <.06, 2-tailed
NA = not applicable